Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9062

Analysis and Prediction of Urban Traffic Accident Correlating Factors Based on SARIMAX Model

Ying Zhu[1.2*]. Wei Yu[1.2]. Hongyan Li[1.2]. Weiyan Yang[1.2] and Yue Lin[1.2]
[1]College of International Business, Zhejiang Yuexiu Univerisity, Shaoxing, 312069, China
[2] Shaoxing Key Laboratory of Intelligent Monitoring and Prevention of Smart City, Shaoxing, 312069, China
Email: 2332702@stu.zyufl.edu.cn

How to cite this paper: Zhu, Y., Yu, W., Li, H. Y., Yang, W. Y., & Lin, Y. (2026). Analysis and Prediction of Urban Traffic Accident Correlating Factors Based on SARIMAX Model. *Journal of Computer Science and Frontier Technologies*, 3(1), 76–85. ISSN Print: 3104-4204; ISSN Online: 3104-4212. https://doi.org/10.63313/JCSFT.9062
Published: 2026-04-20

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Abstract

Nowadays, with the acceleration of urbanization and the rapid increase in population and motor vehicles, frequent traffic accidents pose a threat to public safety, forming the background of this study. Predicting accidents and their resulting losses to provide a basis for traffic management decisions is a key issue to be solved by society. Theoretically, in terms of research significance, the introduction of the SARIMAX model broadens the research path of traffic accidents and promotes interdisciplinary integration. In practice, it can help traffic departments optimize resource allocation and prevent accident risks. Visualized results can provide travel warnings for the public, reducing economic and human losses. The research focuses on urban traffic accident data. It first widely collects multi-dimensional raw data, followed by cleaning, seasonal decomposition, and stationarity testing. Exogenous variables such as weather and holidays are introduced to construct the SARIMAX model. Finally, the Gaode Maps API is used to visualize accident-prone areas. The research methods include literature research, empirical research, and interdisciplinary research methods. The literature method sorts out cutting-edge materials to lay a foundation; the empirical method relies on big data to verify according to the process; and the interdisciplinary method integrates multi-disciplinary expertise to solve problems, providing data support for regional control strategies.

Keywords

SARMAX model; accident data analysis; time series forecasting; spatial visualization; seasonal analysis

1. Introduction

With population growth and accelerated urbanization, road traffic volume is constantly increasing, inevitably raising the probability of traffic accidents. Accidents not only cause casualties and property damage but also adversely affect the normal operation of the traffic system. Therefore, predicting the number and losses from traffic accidents has become an urgent social issue.

Traditional prediction methods are often based on statistical models like regression analysis, which struggle to incorporate the time factor and capture the seasonal and cyclical changes in accidents. Time series methods, built on historical data, are significant for this task. Domestically, research has shown that time series analysis can effectively capture the spatiotemporal distribution patterns of accidents [1, 2]. Internationally, scholars like Amin et al. [3] have successfully used ARIMA models for accident forecasting.

The practical significance of this study lies in aiding traffic policy-making, evaluating safety measures, and supporting public education. The theoretical significance is in filling research gaps, providing a transferable prediction model for macro-situational analysis, and guiding micro-level cause analysis.

2. Literature Review and Methodological Foundations

2.1. Literature Review

Domestic research has applied ARIMA, exponential smoothing, grey prediction (GM(1,1)) [4], and neural networks (BP, CNN, RNN) for accident prediction [5-7]. Internationally, scholars have focused on multi-factor analysis and model optimization, using Bayesian Information Criterion (BIC) for model selection [8], combining time series with machine learning (e.g., SARIMAX-SVM), and applying deep learning like LSTM with attention mechanisms [9, 10].

While various models exist, they often have limitati0ons: ARIMA is weak with complex seasonality, GM(1,1) assumes strong exponential laws and is poor with nonlinear, volatile data. The SARIMAX model addresses these gaps by integrating seasonality, trends, and exogenous variables.

2.2. The SARIMAX Model

The SARIMAX model is a robust time-series analysis tool that extends ARIMA by incorporating exogenous variables, seasonal, and trend components. Its mathematical form is:

$$Y(t) = \mu + \sum \phi(i)Y(t-i) + \sum \theta(j)\varepsilon(t-j) + \beta X(t) + \varepsilon(t) \quad (1)$$

Here, Y(t) is the observed value, X(t) represents exogenous variables (e.g., weather, holidays) with coefficient β, φ and θ are autoregressive and moving average parameters. The model captures self-dependency, lagged errors, and external influences, making it superior for forecasting complex, seasonal time-series data like traffic accidents.

3. Data Acquisition and Preprocessing

The data used in this study comes from public data from insurance companies and a traffic management platform in Zhejiang Province. It contains 120,000 detailed claim records, including structured fields such as accident date, vehicle serial number,

claim amount, number of injuries/fatalities, and detailed accident location, as shown in Table 1.

Table 1. Basic information of the dataset

variable name	Type	Description
Accident date	date	Claim date (format: YYYY-MM-DD)
Vehicle serial number	alphabetic string	Unique vehicle identifier (may include vehicle details such as model and age)
plate number	alphabetic string	Vehicle license plate (partially blank, possibly due to incomplete data or unregistered plates)
Chassis number	alphabetic string	Vehicle identification number (VIN code, used for precise vehicle identification)
Enterprise serial number	alphabetic string	Unique identifier of the involved enterprise (related enterprise information form)
Company full name	alphabetic string	Full name of the company involved (e.g., "Hangzhou No.3 Middle School Municipal Engineering Co., Ltd.")
Enterprise Social Credit Code	alphabetic string	Enterprise unique credit code (for querying registration information and business status)
amount for which loss settled	figure	Insurance payout amount (unit: yuan, ranging from 0 to tens of thousands of yuan)
Number of injured persons	figure	Number of injured persons in the accident
death toll	figure	Number of fatalities in the accident
Accident location details	alphabetic string	Specific location of the incident (e.g., "Jiangqiao Road, Wuyue Street, Lin'an District, Hangzhou")

3.1. Text Preprocessing

Raw text data underwent preparation: 1) Missing values (e.g., license plate numbers) were handled by filling with "Unknown". 2) Redundant fields (e.g., signature values, timestamps) were removed. 3) Date formats were unified to YYYY-MM-DD HH:MM:SS. 4) Outliers in claim amounts (negatives, extreme highs) and casualty figures were detected, and geographic fields were matched against a standard location database.

3.2. Data Visualization

Initial visualizations provided key insights for model building.

Figure 1 is a line chart of the total number of vehicle claims over time. The x-axis is the date, and the y-axis is the claim frequency. It shows significant fluctuations, with a peak near 7,000 claims and drops to near zero on specific dates, suggesting a correlation with special time periods (e.g., February). This indicates a strong seasonal and trend component.

Figure 1. Line chart of total vehicle claim volume changes.

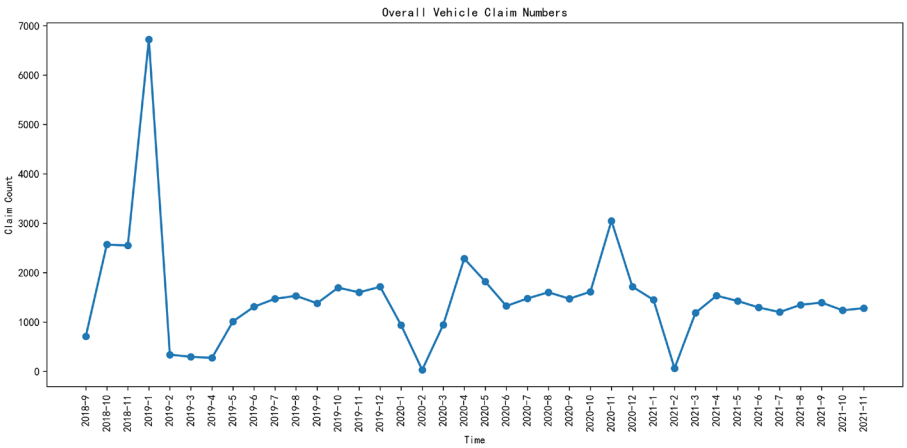


Figure 2. Cumulative distribution of claim amounts.

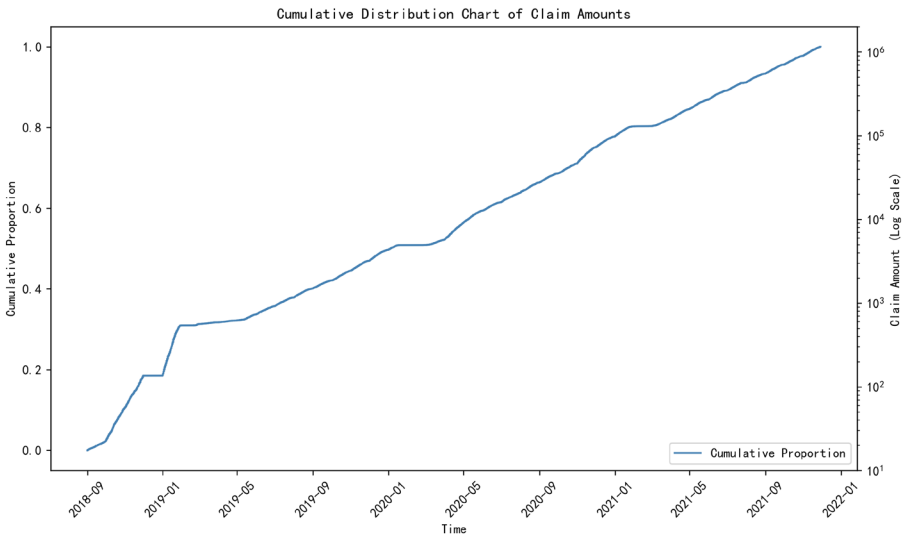


Figure 2 shows the cumulative distribution of claim amounts. The x-axis represents time (Sep 2018 to early 2022) and the y-axis on the left is the cumulative proportion of claims. The blue curve rises steadily, indicating that both the number and total amount of claims have increased over time.

Figure 3. Bar chart of monthly claim volume for 2020.

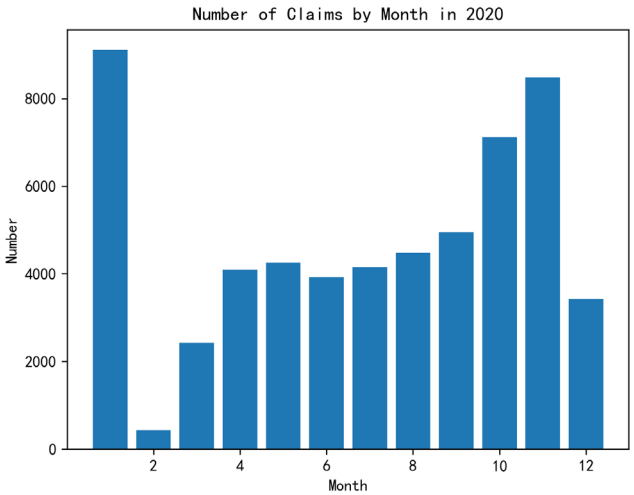


Figure 3 is a bar chart of monthly claims for 2020. The x-axis is the month and the y-axis is the number of claims. It reveals high claim numbers in February (attributed to the Spring Festival holiday and winter weather), a dip afterward, stability from April to June, and a rise from July to October.

Figure 4. Bar chart of total monthly casualties for 2020.

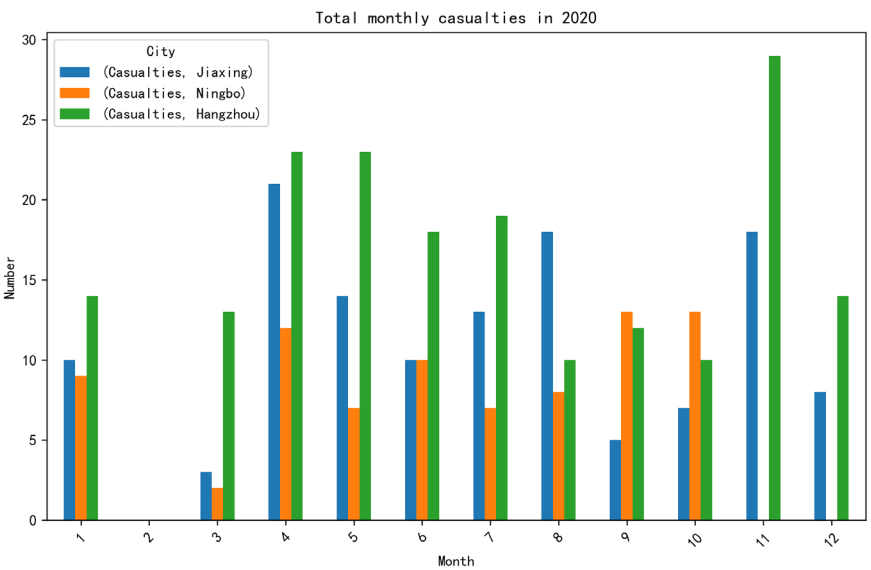


Figure 4 illustrates monthly casualties in 2020 for different cities (Jiaxing, Ningbo, Hangzhou). The x-axis is the month and the y-axis is the total casualties. Hangzhou generally has the highest numbers, likely due to its larger population and traffic volume.

4. Experimental Results

We constructed a SARIMAX model to predict claim amounts and claim volume.

Figure 5. Claim amount prediction chart.

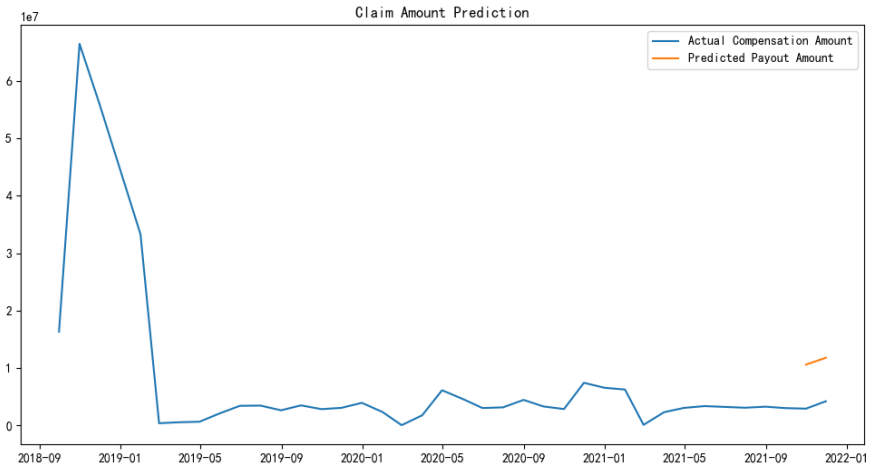


Figure 5 shows the claim amount prediction. The x-axis is the timeline (2018-2022) and the y-axis is the payout amount. The blue line is the ground truth, and the orange line is the predicted value. The model captured the trend effectively, especially the reduced volatility after 2020. The prediction accuracy was high, with an R^2 of 0.86, indicating a strong fit between predicted and actual values.

Figure 6. Claim volume prediction chart.

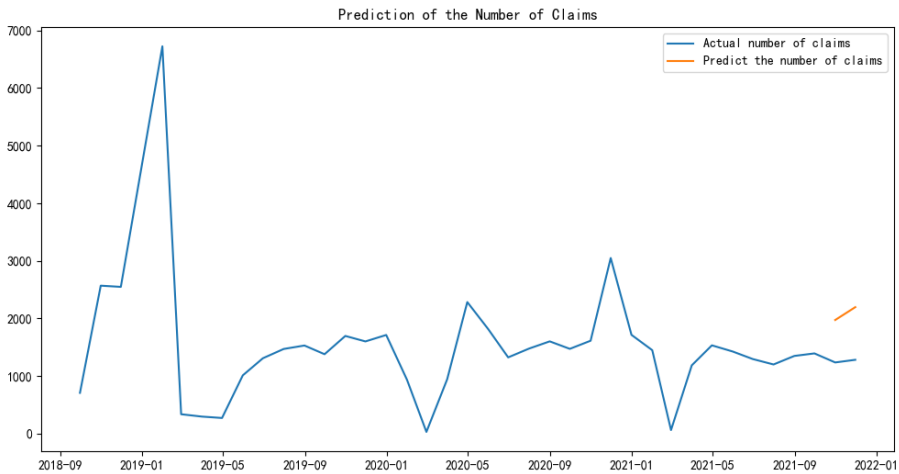


Figure 6 shows the claim volume prediction. The x-axis is the timeline (2018-2022, quarterly) and the y-axis is the number of claims (0-7000). The blue line is actual claims, and the orange line is predicted claims. The SARIMAX model successfully captured the time series trend of traffic accidents in Zhejiang Province.

The R^2 value, calculated by formula (1), is 0.86, demonstrating a high degree of fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

Figure 7 displays the residual plot (x-axis: time 2019-2021, y-axis: residual

values). Residuals were small and stable in certain periods, suggesting a good fit. The residuals also show some structure, indicating the model has captured underlying patterns.

Figure 7. Residual plot.

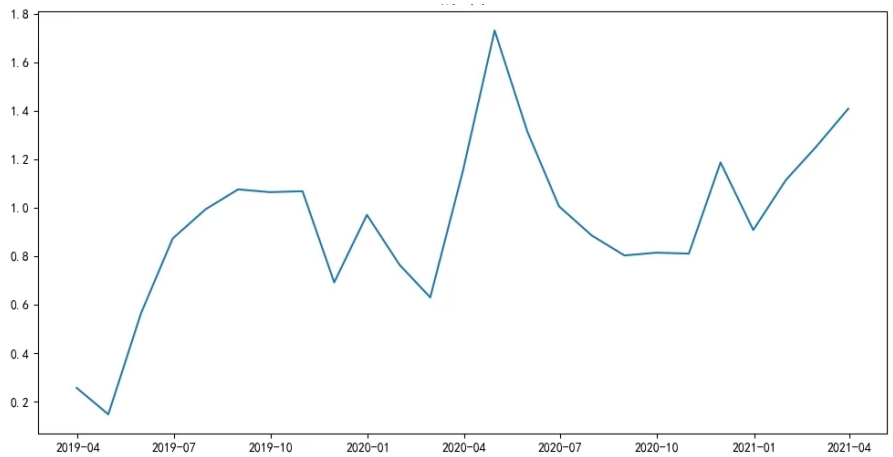


Figure 8 is the fit chart (x-axis: time 2018-2021, y-axis: claim volume). The red dashed line (fitted values) closely follows the blue line (actual values), demonstrating the model's ability to track the overall trend.

Figure 8. Fit chart.

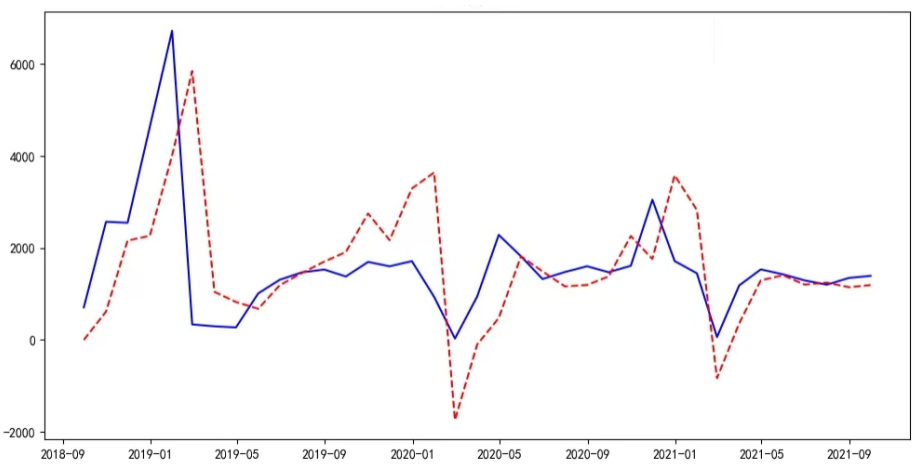


Figure 9. Decomposition plot.

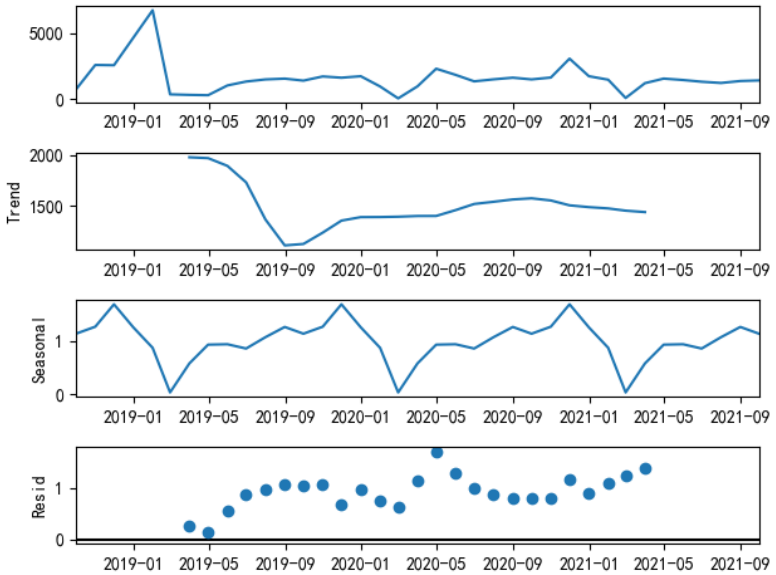


Figure 9 presents a decomposition plot (trend, seasonal, residual components). The trend component shows smooth progression, the seasonal component exhibits regular periodicity, and the residual component fluctuates around zero, validating the model's decomposition.

Figure 10. Diagnostic plot.

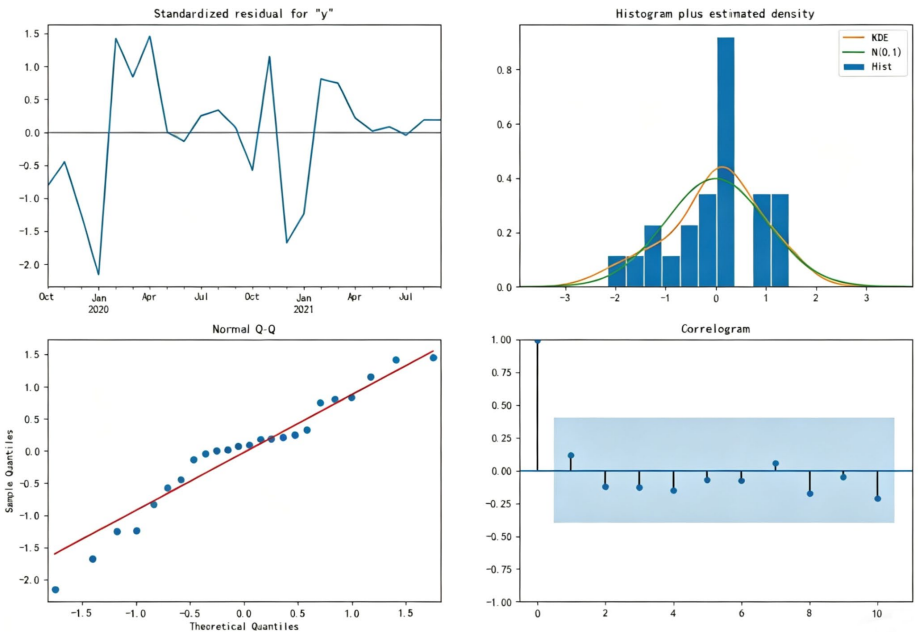


Figure 10. shows diagnostic plots (histogram, Q-Q plot, autocorrelation). The standardized residuals approximate a normal distribution (linear Q-Q plot), and the autocorrelation plot shows most coefficients within the confidence interval,

indicating weak autocorrelation and that the model has effectively separated the data's information.

5. Conclusion

This study analyzed urban traffic accidents using total claim volume and total payout amount as key features. Claim volume directly reflects accident risk and is crucial for traffic management, especially during holidays or bad weather. The total payout amount, related to vehicle damage severity and repair costs, provides a unique perspective for insurance and economic analysis.

The SARIMAX model was introduced for its ability to handle the data's inherent seasonality and autoregressive properties. By integrating exogenous variables, it accurately predicted future accident frequency and payouts. Combined with the Gaode Maps API for spatial analysis, the model provides precise accident location visualization, supporting targeted regional management strategies.

The study has limitations: the data is primarily from Zhejiang Province and focuses on car accidents, limiting generalizability. The SARIMAX model's performance is also sensitive to outliers and data gaps. Future work will expand data collection geographically and by vehicle type, and innovate analytical methods to provide a more comprehensive understanding of global traffic safety.

Acknowledgements

This work was supported by the Zhejiang Province College Students' Science and Technology Innovation Activity Plan (Xinmiao Talent Program) (Grant No. 2025R448A006).

References

[1] Luo, G.C., Gao, J.Y., & Cai, W.X. (2021). Seasonal time series prediction based on GRA-LSTM and SARIMA combination model. Intelligent Computer and Applications, 11(6), 195-200.

[2] Sun, Y.X., Shao, C.F., Ji, X., & Zhu, L. (2014). Traffic accident time series prediction based on ARIMA and information granulation SVR combination model. Journal of Beijing Jiaotong University, 54(3), 348-353.

[3] Amin, S.M., et al. (2007). Road accident forecasting models using time series analysis and artificial neural networks. Accident Analysis & Prevention, 39(2), 311-320.

[4] Wang, Y.J., & Sun, Y.X. (2007). Road traffic accident prediction based on the GM(1,1) model. Traffic Technology and Economy, 9(5), 97-98.

[5] Liu, Z.Q., Shan, X.M., Xin, N., et al. (2024). Research on traffic accident prediction based on ES-ARIMA-BP neural network with combined weighting optimization. Modern Electronic Technology, 47(22), 71-76.

[6] Yang, X.M. (2000). Fuzzy neural network prediction model for bus route passenger flow. Highway Traffic Technology, 17(4), 38-41.

[7] Qiao, Y.Q. (2022). Traffic Accident Prediction Based on Deep Learning (Master's thesis). Xi'an University of Science and Technology.

[8] Zhao, J.B., Deng, W., & Wang, J. (2011). Urban road traffic accident analysis based on Bayesian network. Journal of Southeast University (Natural Science Edition), 41(6),

1300-1306.

[9] Zhang, J., Liu, X.M., He, Y.L., et al. (2007). Application of the ARIMA model in traffic accident prediction. Journal of Beijing University of Technology, 33(12), 1295-1299.

[10] Wu, Y.D. (2023). Traffic Accident Prediction and Analysis Based on an Improved LSTM Neural Network Model (Master's thesis). Chongqing Jiaotong University.